Class Maker 1.2.7 ß

User Manual

Gianluigi Cardinali

University of Perugia
Dipartimento di Biologia Vegetale
Sezione Microbiologia Applicata

Laboratory of Molecular Genetics and Evolution

Borgo 20 Giugno, 74 - I-06121 Perugia - Italy
Tel +39 075 585 64 84 - Fax +39 075 585 6470
e-mail : gianlu@unipg.it

Contents

		Pag.
1	Background a. Practical Problems in banding pattern analysis b. Scientific Problems in banding pattern analysis	2
2	ClassMaker as a linking tool with other free software	
3	Calculating migration distances of bands with NIH-Image	
4	Calculating Molecular weights with MacCurve Fit	
5	Class Maker: operating functions a. Start b. Define Samples c. General considerations on band classification d. CL1 e. CL2 f. CL3 g. Classify	
6	Processes on binary matrices	
7	Statistical analysis with "Le Progeciel"	
8	Statistical and Phylogenetic analysis with "Phylip"	
9	Combined use of Le Progeciel , Phylip and tree drawing programs	
10	Phylogenetic analysis with PAUP and MacClade	
11	Ad maiora	

A. Background

The statistical and phylogenetic analysis of complex banding patterns obtained with various molecular techniques is one of the most powerful tools of the biologist for the characterization (or typing or fingerprinting) of organisms and populations, for diagnostics, forensic tests and other applications. Although quite a few algorithms, analytical approaches and free software have been designed to analyze characterization data, the number of statistical analyses routinely used in molecular biology laboratories is relatively small. In fact most of the dedicated commercial packages use only the UPGMA algorithm or a few others. The gap between the available analytical procedures and those actually used is not apparently decreasing with the availability of more powerful packages, mostly designed to reduce the weight of the operator on the whole analytical process.

ClassMaker is free software usable on both Mac and PC Computers, designed to link free domain packages capable of performing one or a few steps of the whole analysis of the banding patterns and to classify the bands according to non-subjective automatic routines.

1.a Practical Problems

After a long series of molecular procedures you have your gel on the transilluminator and a picture of it. Now the question is : -what next?-.

Next comes an appropriate statistical or phylogenetic analysis of the banding patterns one has obtained. But, once again, which analysis? One should consider statistical analysis when the focus of the research is the similarity of the organisms under study. This analysis has conceptually three phases: 1. Transform your bands in a matrix of data, 2. assess the statistical distance among the strains (or individuals) and 3. construct a tree (or dendrogram). Alternatively, one can use the original matrix (point 1.) or the distance matrix (point 2.) for other analyses such as the Principal Component Analysis (PCA) or the Principal Coordinates Analysis PCoA (for references see: http://www.fas.umontreal.ca/BIOL/legendre/indexEnglish.html).

Dendrograms are normally more difficult to read but take into consideration the whole variability of the population, while PCA or PCoA are easier to read but express only a part of the whole variability as reported on the axes of the graph.

Let's take for instance a matrix of 8 strains each described by seven characters (Tab. 1) and let's follow the possible treatments outlined above. The matrix of Tab. 2 represents the distances among the eight strains calculated as normalized Euclidean matrix. In Fig. 1 there is displayed a dendrogram reconstructed with the Neighbor Joining algorithm, whereas in Fig. 2 and in Fig. 3 there are reported the results from the PCA and the PcoA. In general one must take care to perform PCA with quantitative (not classified data), whereas PcoA can be carried out with any type of data. In the case shown in Fig. 2 and Fig.3 it is clear that both algorithms produce the same overall topology of strains, indicating that even classified data could be processed with PCA. In any case the availability of both algorithms in the same package (Le Progeciel) allows the choice of the appropriate one.

Programs as Le Progeciel and PAUP allow several statistical and phylogenetic analyses of initial data such as those presented in the matrix of Fig. 1. Otherwise the problem is to obtain such a matrix starting from the bands of the gel. ClassMaker has been expressly designed to fill this gap and allow investigators to produce their matrices in a matrices way in order to access all possible types of statistical and phylogenetic analysis.

Tab. 1 : matrix of data in binary form (0= character absent; 1= character present), describing 8 strains with seven characters.

	char1	**char2**	**char3**	**char4**	**char5**	**char6**	**char7**
Strain A	1	1	1	1	0	0	0
Strain B	1	1	1	1	1	0	0
Strain C	1	1	1	1	0	0	0
Strain D	0	0	0	0	1	1	1
Strain E	0	1	0	1	0	1	0
Strain F	1	0	1	0	1	0	1
Strain G	1	1	0	0	1	1	0
Strain H	0	0	0	0	1	1	1

Tab. 2 : distance matrix derived from matrix of Tab. 1 using the program Le Progeciel with the distance algorithm D2. Note that the two triangular matrices separated by the diagonal are exactly symmetric. Distances among identical strains are obviously = 0 as visible in the descending diagonal and in the four (two pairs) other value that indicate that A is identical to C and H to D. On the other hand, totally different strains (no matches in the character matrix) have a distance of 1 (e.g. A and D).

	Strain A	**Strain B**	**Strain C**	**Strain D**	**Strain E**	**Strain F**	**Strain G**	**Strain H**
Strain A	**0.000**	0.378	**0.000**	1.000	0.655	0.756	0.756	1.000
Strain B	0.378	**0.000**	0.378	0.926	0.756	0.655	0.655	0.926
Strain C	**0.000**	0.378	**0.000**	1.000	0.655	0.756	0.756	1.000
Strain D	1.000	0.926	1.000	**0.000**	0.756	0.655	0.655	**0.000**
Strain E	0.655	0.756	0.655	0.756	**0.000**	1.000	0.655	0.756
Strain F	0.756	0.655	0.756	0.655	1.000	**0.000**	0.756	0.655
Strain G	0.756	0.655	0.756	0.655	0.655	0.756	**0.000**	0.655
Strain H	1.000	0.926	1.000	**0.000**	0.756	0.655	0.655	**0.000**

Fig. 1 : Neighbor Joining dendrogram obtained from the distance matrix of Tab. 2 with the T-REX program

Note that identical strains (A and C; D and H) share the tip of the same branch

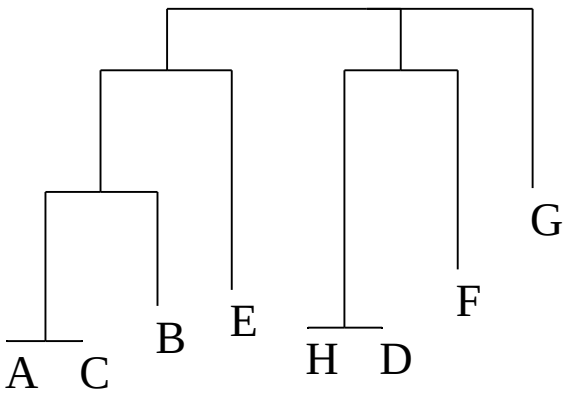


Fig. 2: PCA scatterplot obtained from the matrix of Tab. 1 with Le Progeciel .
The variance displayed is 85%: 59% on the X and 26% on the Y axis.

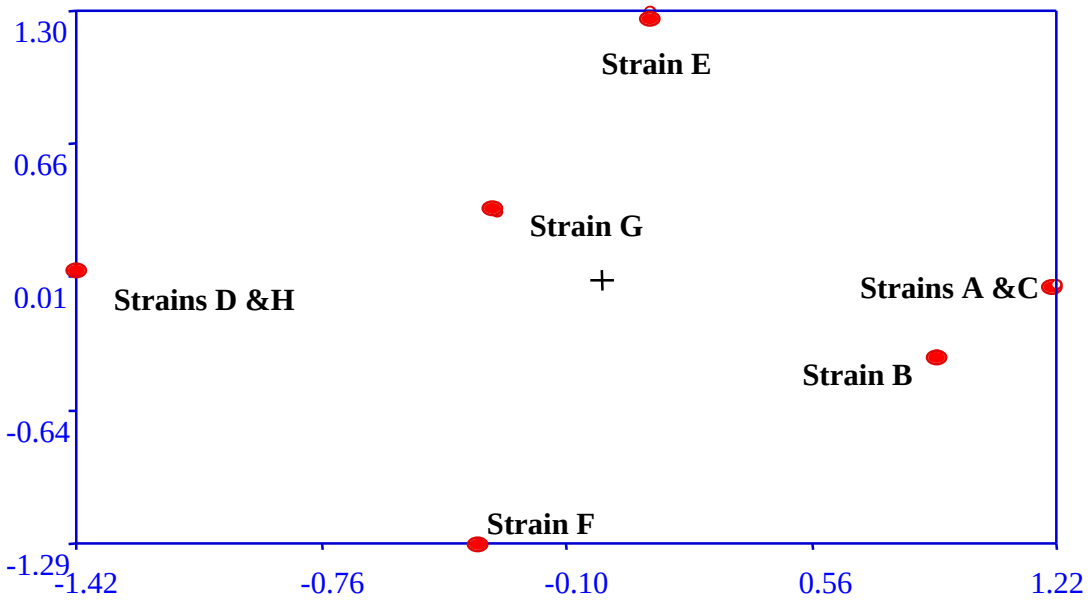
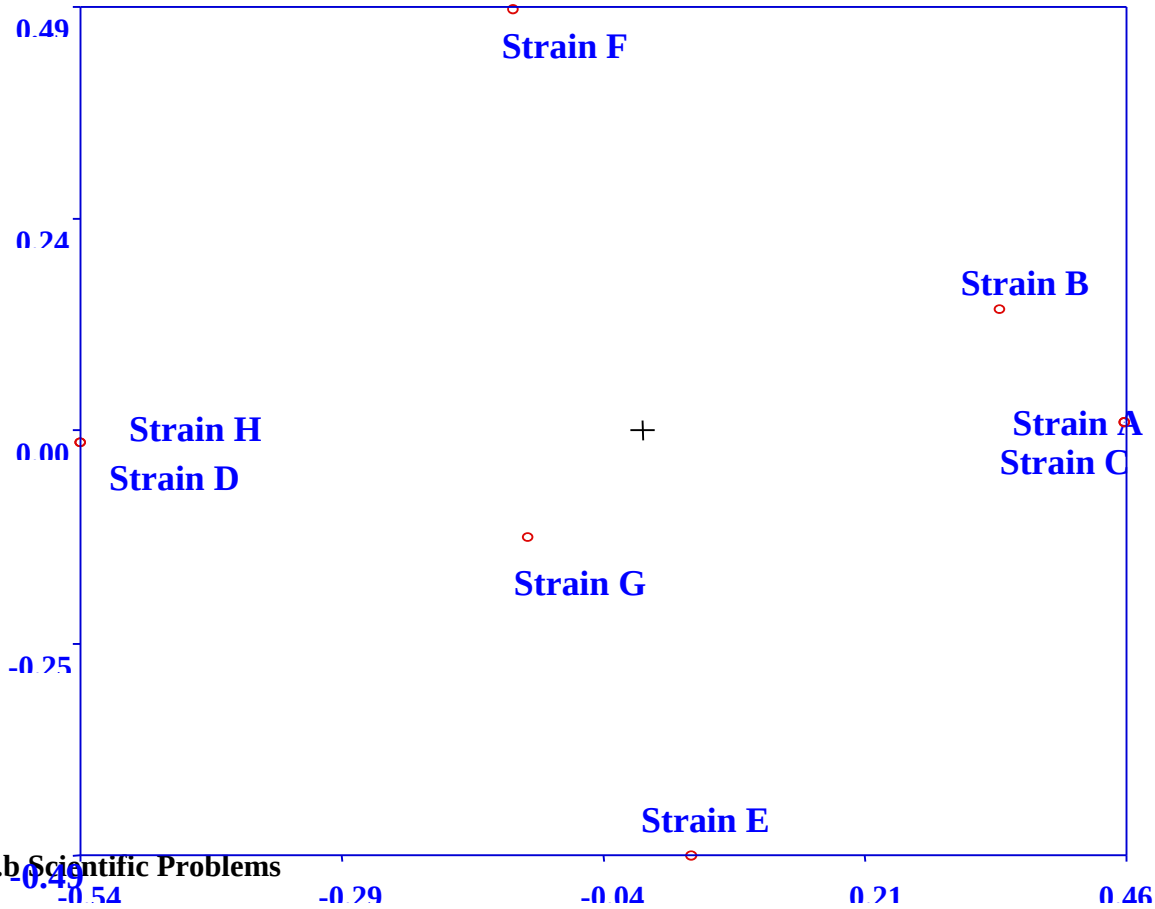


Fig. 3: PCoA obtained from the distance matrix of Tab. 2 with Le Progeciel. Note that the overall topology is similar to that of Fig. 2 with a rotation around the horizontal median axis. The variance displayed is 85%: 59% on the X and 26% on the Y axis, as above in the PCA of Fig. 2.



1.b Scientific Problems

One of the main problems the investigator normally has to cope with is choosing the most appropriate type of analysis for the data available, in order to solve the particular problem under study. As a matter of fact not all algorithms are appropriate for all investigations, depending on the biological model and on the type of data. Most dedicated commercial software uses UPGMA (which assumes a constant mutational rate) to construct dendrograms. Unfortunately, the mutational rate of many subjects (genes, proteins, organisms) is unknown. For an extended review on the topic refer to Hillis [Hillis, 1996 #3406] . Altogether, these considerations suggest that the system of analysis should be flexible enough to be used with all available statistical and phylogenetic algorithms, in order to match the precise requirements of the investigator.

B. ClassMaker as a linking tool with other free software packages

When considering the whole analysis from the gel to the final output, a discouraging consideration is the necessity of using several software programs with different input and output formats. Whereas most of these incompatibilities can be overcome with easy changes of format, the real gap in the series of operations from gels to dendrogram is the fact that image analysis software, as NIH-Image, gives as output a column of quantitative value of band migration distance. On the other hand, most of the programs useful for further statistical or phylogenetic analyses require as input a matrix of data organized with the objects (the strains or species) in rows and the variables (the characters analyzed) in columns (see Tab. 1). Moreover, some statistical and almost all phylogenetic treatments require binary matrices where 1 and 0 mean, respectively, presence and absence of the character. Filling this gap requires the following operations:

A. transforming migration distances into molecular weights
B. transforming the column of migration distance data into a matrix corresponding to the strains (the different lanes of the original gel) and to the characters (the bands in the gel).
C. producing a non-subjective classification system of the derived molecular weights
D. automatically classifying the bands into classes of molecular weight produced in the former step.

Operation **A**. can be carried out with several free or commercial software programs that construct a calibration curve using the migration distances and the corresponding molecular weights of the DNA marker included in each gel. The molecular weight of each band is calculated according to the regression equation. From a practical point of view this means that a column of distance data is transformed into a column of molecular weight values.

ClassMaker has been designed to carry out the other three operations, producing a binary matrix arranged according to the strains and the characters (see Tab. 1), which can be further processed.

In the next chapters there will be described the whole series of operations in which free software can be used to carry out a complete analysis of a gel, starting with the determination of the migration distances of the bands of a gel with NIH-Image (Chapter 3). Then, in Chapter 4, the system for calculating the molecular weight of each band with MacCurve Fit will be illustrated . In Chapter 5 ClassMaker functions will be described in detail along with the changes of format necessary to use the binary matrix as input in the various programs of phylogenetic or statistical analysis.

One of the most important features of Class Maker is that matrices from different analyses (e.g. different RAPDs, RFLP, AFLPs etc) can be merged to form a large comprehensive matrix in which the strains are described by the variables obtained from all analyses. Finally, in Chapters 7, 8,9 suggestions will be made for analyzing matrix data with some of the best known free or commercial packages for statistics or phylogenetics.

C. Calculating migration distances of bands with NIH-Image

Picture digitalization

The first step in processing a gel is to get a digital picture. This can be achieved either by scanning a hard copy (e.g. a Polaroid instant picture) or by capturing the gel image with a TV camera connected to any video card. In our case we used a B/W TV camera (Kappa GmbH - Germany) connected with a Nu-Vista card on an Apple Quadra computer. Images were captured and saved in TIFF format. Any other combination of camera, computer and video card already present in the investigator's lab can be used: the only important thing is to obtain a digital picture.

Picture processing

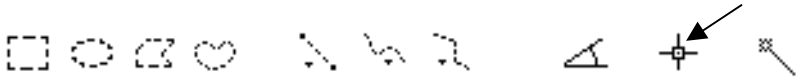
The rule of thumb in picture processing is that the less you do the better, because any operation can produce artifacts. However, there are pictures requiring some general operation, such as increasing or decreasing luminosity, contrast, etc. Another necessary operation is correcting deformations in the gel such as the **smile effect** visible in those gels where the central lanes migrated more that the external ones, or **tilted gels** where the lanes of one side migrated less than the those of the other side. In this case one can try to tilt the picture in order to obtain a correct positioning of bands, if identical lanes (i.e. identical DNAs in different lanes) are present in the different areas of the gel. In our experience a smiling gel should be discarded and remade, whereas tilted gels can be corrected if the same DNA marker was loaded in the two most external wells. In this case we distort the image using Adobe Photoshop in order to have the same bands of the molecular standard migrating the same distance from the gels. However, it is very important to consider that no operation should be undertaken on single lanes or bands, but only on the whole gel, in order to avoid specific changes affecting only some elements (bands or lanes) of the gel. It is important to remember that no analytical system can help if the original picture is too bad.

Migration distances measurements with NIH- Image

The gel image should be opened with NIH-Image (we used version 1.62 for Mac, but any other is fine) and inverted (Menu Edit: Invert). The NIH-Image Manual recommends subtracting the background (Menu Process, subtract background) with the "2D rolling ball" system. Since bands tend to fade when this is done, one should take care that the least visible do not disappear. There are basically two systems for measuring band position and migration distance, hereinafter referred to as md: one uses the cross hair tool (see Fig. 4), the other the "Gel Plotting macro".

Fig. 4 Tool palette of NIH-Image

cross hair tool

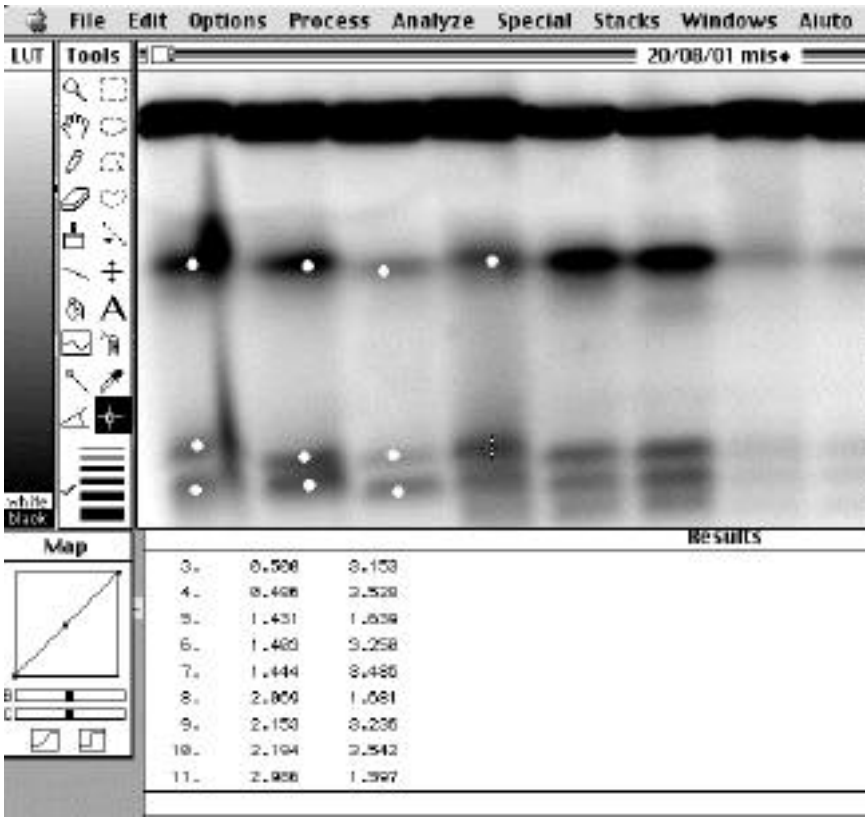


Measurements with the cross hair tool

Open the **Show results** window from the **Analyze** menu and place aside (or below) the window with the gel (Fig. 5). Open **options** from the **Analyze** menu and check only **X-Y center**, set decimals at three digits. Position the cross hair tool on the center of each band and click on it. Check that each new measurement has been recorded in the **Show results** window.

From the **Edit** menu choose **Copy measurements** and paste the data into a new window of Mac Curve Fit, or any other program able to calculate regressions.

Fig. 5 Measuring migration distances with the cross hair tool of NIH-Image



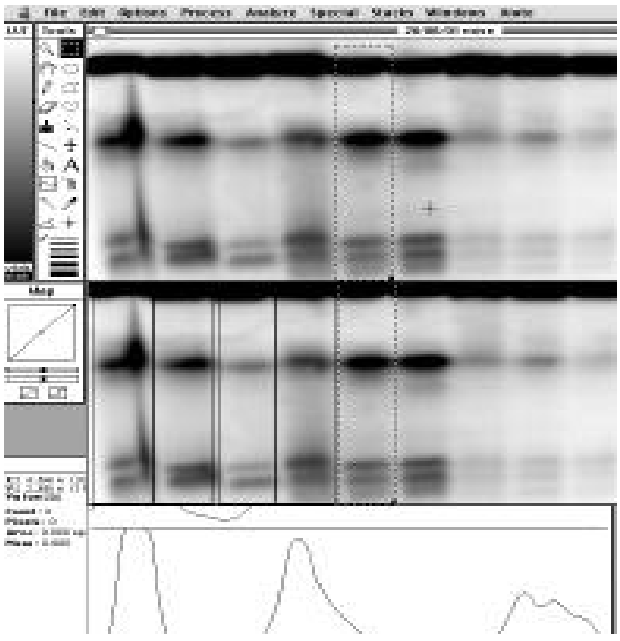
Measurements with the Gel Plotting Macro

Open the **Special** menu and click on **load macros,** choosing the **Gel plotting** macro from the macros folder of NIH-Image . Identify the first (the left-most) lane and from the **Special** menu choose **Mark first lane** (or alternatively, type "1"). Another window displaying the gel will appear; work on this window. Now identify the second lane (in the new window) and choose **Mark next lane** (or alternatively, type "2"); repeat this operation for all lanes. After the last lane has been marked, choose **Plot lanes** (or alternatively, type "3"). A window will appear with densitograms of all lanes (Fig. 6). Open the **Show results** window from the **Analyze** menu and place it aside (or below) the window with the gel (Fig. 5). Open **options** from the **Analyze** menu and check only **X-Y center**, setting decimals at three digits. Position the cross hair tool (or the wand- tool) on the top of each peak and click on it. check that each new measurement has been recorded in the show results window.

From the **Edit** menu choose **Copy measurements** and paste the measurements into a new window of Mac Curve Fit, or any other program able to calculate regressions.

For all details on NIH-Image, please refer to the manual, freely downloadable along with the program from the NIH site.

Fig. 6 : Measurements of migration distances with the "gel plotting macro" of NIH - Image



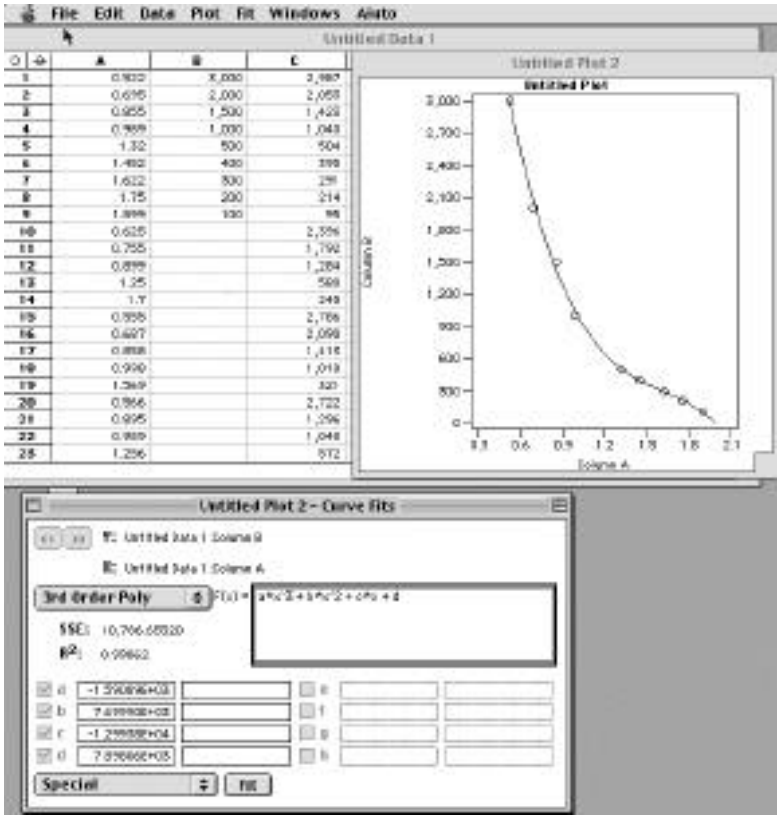
D. Calculating Molecular weights with MacCurve Fit

Open a new window and paste into it the column of results from NIH Image.

If the measurements were carried out with the hair cross tool the column A will report the horizontal position of the bands (X-axis) and the B column will display the vertical position, corresponding to the migration distance (Y-axis). Eliminate data in column A and move onto it the data of column B.

As indicated in Fig. 7, column A will now contain the m.d. of each band (X-axis).At this point the B column will contain the peak heights (Y-Axis). Eliminate this column B and enter there the known molecular weights of the ladder (Fig. 7). You are now ready to plot the standard curve.

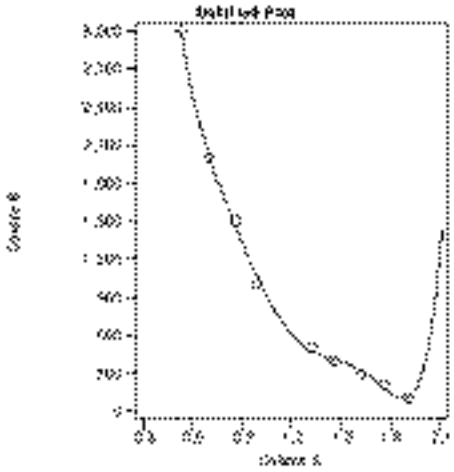
Fig. 7 : The three windows of Mac Curve Fit involved in calculating the molecular weights from the m.d. of the bands.



From the **data** menu**,** choose **plot data;** a window will appear with a scattergram. From the **fit** menu**,** click **curve fit** and then choose one of the several regression options. In our experience

polynomial regression curves of odd degree give the best results. It is necessary to check that the curve does not show plateaus or ditonic areas where different m.d. values correspond to the same molecular weight (see the right part of the graph in Fig. 8). In such cases the best option is to reduce the degree of the polynomial function.

Fig. 8 : Regression curve with a final ditonic area



Click on **Fit**, a regression curve will appear on the scattergram, while on the fit window the R^2 and the SSE values will be displayed (Fig. 7). In our laboratory only regressions with $R^2 \geq 0.99$ are accepted.

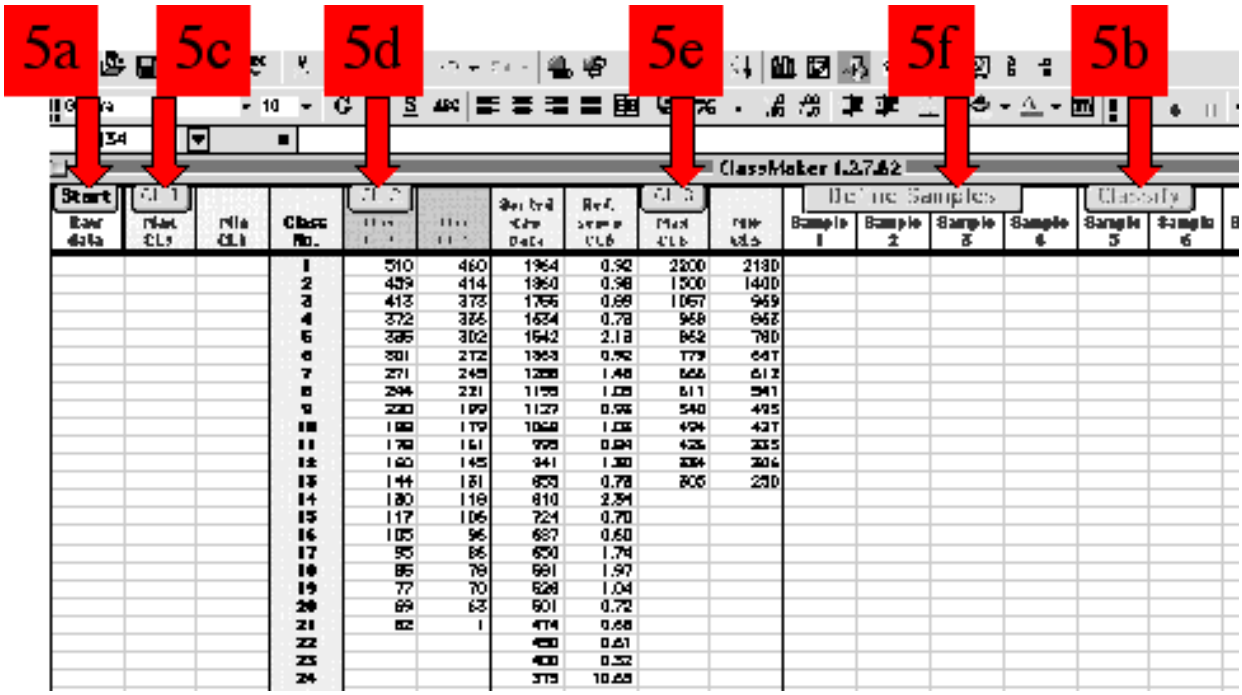
Finally, from the **Fit** menu choose **predict batch Y** and set column C as destination of the predicted Y values (Fig. 7). Transfer molecular weights of all bands from the C column of MacCurveFit into the first column of Excel and save this file. Start the ClassMaker macro and import data from the previously saved file with the **start** function of ClassMaker. It is possible to combine data from more gels, obtained with the same molecular procedure, by simply adding the data to the column.

E. **Class Maker: operating functions**

ClassMaker is a free Excel™ macro written to run in both Mac-OS, or WIN platforms. It has been designed to transform a column of quantitative data into a matrix of discrete (1/0) values, from now on referred to as binary matrix, classified according to one of the three criteria offered by the macro. Six functions are implemented**: Start, Define samples, CL1, CL2, CL3, Classify**. It can be freely downloaded from http://www.agr.unipg.it/cardinali/index.html.

First of all a glance at the ClassMaker window.In Fig. 9 the window is displayed with all six functions and the sections in which they are described.

Fig 9 : The window of ClassMaker with its six functions. Numbers over the arrows indicating the functions refer to the sections of the text.



5a. Start

This function allows the importation of a column of quantitative data into the first column (A) of the working sheet of ClassMaker. A pop out window indicates the number molecular weight data imported (labeled as **row data** and corresponding to the total number of bands). After this the number of row data remains in the memory of the program until the window is closed. The column of row data, coming from the calculation with the regression equation, can be introduced in Class Maker either by a copy / paste procedure or by clicking on start. The latter operation should be preferred because it stores in memory the number of row data, which is an essential input for all classification functions.

5b. Define Samples

This function transforms a column of data into a matrix. Programs such as NIH-Image store coordinates of single bands (Y and X values) in a single column without differentiating between lanes or samples. The "define sample" function has been designed to reconstruct the distribution of data according to the strains, starting from a single column of data such as that obtained with NIH-Image.

5 c. General considerations on band classification

Since it is difficult to choose *a priori* the right system of classification, the following three intuitive criteria can be used to accept or reject possible classifications:

- The number of classes must be at least as large as the number of bands in the sample with the most bands.
- The matrix obtained with the classification must be binary and therefore contain only 0 and 1 values.
- Two independent replicates of the same sample must have the same pattern of classification. This is normally obtained by including a sample twice in each experiment; in the case of electrophoretic gels, it is advisable that replicates run in distant lanes.

For a classification to be acceptable, all three criteria must be satisfied. Sometimes it can be hard to meet the first criterion, for instance because there are two very close bands falling into the same class. The problem can be solved in at least two ways: adding an intermediate class to those produced by the program, or accepting a trinomial matrix.

Classification systems **CL1, CL2, CL3**.

These three functions produce two columns containing the upper and lower limits of the classes (called Max and min, respectively) in decreasing order. Data in the columns will be referred to as M1, M2….Mn, and m1, m2…mn for the series of Max and min values respectively. In all three systems, the second column (min) is automatically obtained applying the formula

(formula 1) $m_i = M_{(i+1)} + 1.$

The classification system to employ can be selected empirically, taking into consideration the three criteria described above. From a practical point of view it is important to point out that the operator can produce classifications with one of the **CL** algorithms, without disturbing the others. Moreover, one can classify a matrix of quantitative data with several different combinations of algorithms and parameters; this is possible by choosing the "**starting row**" in the last dialog window of **Classify**. For simplicity, algorithms and parameters will be indicated hereinafter in a short fashion, for instance "**CL3-75e**", indicating that the **CL3** classification system has been used, with 75% of Similarity and the **e** algorithm.

The practical approach to find the best classification system is to enter extreme settings and then to adjust them according to the classification results. For instance, if **CL3-0c** gives too few classes , one can immediately try a **CL3 -100 c**. If there are now too many bands, one starts looking at **CL3 -50c** and so on.

5d. CL1

The rationale of this classification system is that different classes contain bands differing by a threshold value expressed as percentage, implying that the absolute value of the minimum difference necessary to create a new class is proportional to the Mol. Wt. of the bands. This treatment is in accord with the fact that the best resolution is obtained in the lower part of the gel where the lighter bands migrate, so that bands of lower Mol. Wt. can be discriminated more finely than those of larger Mol. Wt. Formally, the concept is expressed by the formula (2), where T represents the threshold value:

(formula 2) $M_{(i+1)} \leq M_i \times T\%$

When the **CL1** function is activated, the operator is asked to enter a T% value, then the macro produces two columns: **CL1 Max** and **CL1 min,** ordering data in decreasing order in the two columns dedicated to **CL1**. Maximum values are defined according to

formula 2, starting from the first (largest) value. The series of min values is calculated as described above (formula 1).
It is important to point out that the Max values are actual Mol. Wt. data obtained from the gel and not artificially constructed class limits such as those produced in the **CL2** function.

Practical considerations

In **CL1** the operator introduces only the value of threshold which represents the minimum difference between the upper (or lower) limits of two contiguous classes. If the **T%** value chosen is too small classes will be too broad, producing classes which contain more than one band. On the other hand, if the **T%** value chosen is too high, one risks placing homologous bands in different classes.

5e. CL2

In this algorithm, presented as formula 3, The investigator must enter the class amplitude and then choose number of classes desired. This algorithm calculates Max values, applying the formula 3, starting from the highest Mol. Wt. value of the first column of row data.

(formula 3) $M_{(i+1)} = M_i \times A\%$.

In formula 3, **A%** is the class amplitude defined as the ratio between two contiguous Max values multiplied by 100 and therefore expressed as percentage. As the system is independent from the actual row data, it is necessary to enter the number of classes one wants to produce, in addition to **A%.**

Practical considerations

The **Class Amplitude (A%)** parameter of **CL2** is the complement of **T%** in **CL1;** the main difference between the two systems is that **CL2** produces homogeneous classes without considering the actual values present, while **CL1** calculates from the Mol. Wt. values.
The number of classes desirable should be estimated on the basis of the first criterion of classification described above. In practical terms, the investigator should have run **CL1** in advance, in order to know the lowest Mol. Wt. value within the row data (ie. the last Max value). Now, the number of classes in **CL2** is chosen in such a way that the last class of **CL2** has a Max value not larger than the Max value in the last class of **CL1**. Usually one asks the program to produce a number of classes which is 2 or 3 times larger than

the highest number of bands found in the lanes of the gel. If the last Max value is too small, then the investigator should remove those classes with Max values lower than the Max value of the last **CL1** class. If the last Max value is too large, then more classes are needed.

5f. CL3

This function calculates classes with the following series of operations. First, it sorts row data in decreasing order in column **sorted row data (rd)**, using the equation of formula 2. Next it calculates a score of similarity between values of the first column according to one of the seven formulas explained below and writes them in the column **Reference score of CL3**. Finally the program sets the upper limits (Max) of each class according to formula 4

(formula 4) $M_i = RS_{(i-1)} > 1$

where **RS** is the reference score calculated in the second step by one of the seven available algorithms.

The input required from the investigator is first, the choice of a **Band Similarity Threshold** (**BST**) as described in "practical consideration" section and then the choice of the algorithm marked with letters from **a** to **h**.

DST values, ranging from 1 to 100, correspond to the range 90 to 100 of **T%.** The relationship between **T%** and **DST%** is **T%=0.9 + (DST x 10**). This makes **DST** values more sensitive than those of **T%**.

In the second step the " **reference score of CL3**" is calculated using one of seven algorithms offered by the macro. They are arranged in such a way that **a** produces fewer classes while the others yield more classes up to the maximum number, usually obtained using **h**.

The point of **CL3** is the detection of discontinuities. These can be found by comparing the differences between each single datum and those preceding and succeeding it. Of course, the effects of this comparison will vary depending on the range of data considered before and after each single value. The range of data involved in the calculation can be chosen by trying the formulas of each algorithm, reported in the following equations, in which sorted molecular weight data are reported as D1, D2, D3….Dn with D1>D2>D3>…>Dn (where D is a datum).

a. D2-D3 / D3-D5
b. 2 x (D2-D3 / D3-D5)
c. D2-D4 / D3-D5
d. D2-D3 / D3-D4
e. D1-D4 / D2-D5
f. D1-D3 / D3-D4
g. D1-D3 / D2-D5.
h. D2-D3 / D2-D1

In this algorithm each datum becomes in turn D1, D2, D3 etc. Each calculated **RS** is reported in the column **Reference Score of CL3**. This means that when **RS** > 1, the figure corresponding to D3 is taken as the upper limit of a class. The program reiterates this procedure until the end of the row data.
In **CL1** and **CL3** the last class has the **Max** value corresponding to the smallest value within the row data and "1" as **min**.

Practical considerations

In **CL3** the operator must introduce the **Band Similarity Threshold** (**BST**) which resembles T% (it ranges from 0 to 100), but is very different because **BST** ranges only from 90 to 100%. In fact an input as **BST%** = 50 is equal to **T%** = 95. The reason for this setting is that **CL3** is much more sensitive than **CL1** and requires a finer tuning of the threshold. Note that in both in **T%** and **BST**%, the input of 100 will only subtract precise duplications of values from the column of the sorted raw values. However, in most cases, especially those with many bands per lane, a high **BST%** value is not desirable. Such high settings would introduce artifacts because homologous bands may occupy slightly different positions. If the **BST%** value is too low then non-homologous bands will occupy the same class. Altogether this means that ClassMaker offers an automatic classification system, but asks the investigator to judge each classification in the light of the three criteria described above.
The choice of the algorithm (**a** to **h**) in **CL3**, mostly depends on the distribution of bands in the gel. Matrices with many bands in the same lane, differing by only a few bp will require more classes, usually produced by the algorithms after **c**. Very "compact" matrices can require the algorithm **h** as indicated in the dialog window. The algorithm **c** has been designated as "default" because it produces an intermediate number of classes.

It must be noted that succeding algorithms will not always produce an increasing number of classes depending on the specific distribution of the bands.

A basic requirement for the application of the three classification criteria, illustrated above, is to have at least two replicates of the same sample. This can be difficult or impossible in some experimental situations, e.g. when not enough wells are available. However in such cases it is possible to copy one of the left-most lanes in the right-most side of the gel, obtaining a virtual replicate that can be used to apply the third criterion of classification. Other procedures are now under study in our laboratory.

5g. Classify

The **Classify** function assigns the bands of each sample (expressed as Mol. Wt row data.) to the classes defined with one of the three systems described above. The operator is requested to choose the classification system, by entering 1, 2 or 3 for **CL1**, **CL2** or **CL3**, respectively. The investigator now has to enter the number of classes to employ and the number of samples to be classified. Finally, the program asks from which line the classified figures (1/0) should be introduced. Of course, the operator will chose a line well below the original matrix of quantitative data. The investigator can choose each classification system in turn entering the new matrix below the previous ones and then decide which is preferable according to the three criteria of classification. Overwriting is not allowed, because the program would sum the figures of the overwriting matrix with those of the previous one.

Final matrices can be saved in a new Excel file for the further steps of the analysis. Note that the last row of the binomial matrix contains only "0"s to indicate that the classification has been completed; this row should be deleted.

E. **Processing binomial matrices**

Checking the binary matrix.

In order to check which of the possible matrices is correct and congruent with the original data, we suggest a simple test based on the following three intuitive criteria:

1. The number of classes must be at least as large as the number of bands in the sample with most bands. If one has not recorded the number of bands of each strain, it is possible to activate the function **Define samples** immediately upon importing the data and before

starting to search for the optimum classification. The matrix obtained allows an easy visualization of the bands in each sample.

2. The binary matrix obtained upon classification must contain only 0 and 1 values.

3. Two independent replicates of the same sample must have the same classification pattern. This is normally obtained by including a sample twice in each experiment; in the case of electrophoretic gels, it is advisable that replicates run in distant lanes.

For a classification to be acceptable, all three criteria must be satisfied. Sometimes it can be hard to meet the second criterion, for instance because there are two very close bands falling into the same class. The problem can be solved in at least two ways: by manually inserting an intermediate class to those produced by the program, or accept a trinomial matrix. The latter solution can only be used with programs able to manage figures different from 0 and 1, such as MacClade; however, one must carefully check that the programs employing this classification do not produce artificacts when values higher than 1 are used as input.

The test for criterion 1 is carried out when choosing the classification system (**CL1, CL2** or **CL3**); criteria 2 and 3 are checked after the binary matrix has been written by ClassMaker. The simplest system to test the presence of figures other than 0 and 1 is to use the **find** function of Excel.

Testing criterion 3 can be difficult in large matrices because the columns being checked may be very distant and it may happen that one of them is out of the window. A good solution in this case is to copy both columns side by side in another part of the worksheet. An upgrade of ClassMaker, which will perform this test automatically is planned.

Criteria 2 and 3 are clearly complementary, so that very broad classes will produce identical patterns. Unfortunately, such classes may yield a matrix containing “2”s or “3”s. Conversely, narrow classes will never produce values other than 0 and 1, but identical bands might be assigned to adjacent classes, thus producing an artifact. The above considerations explain why a binary matrix can be accepted only if all the three criteria are satisfied.

Preparing the binary matrix as input for further applications.

A procedure for reducing recognized software incompatibilities is to copy the matrix in a new file using the **copy special** command of the **Edit** menu of Excel and checking the boxes "**transpose**"

and "**only values**". The new file should be saved in **text delimited by tab** (or by spaces) format. These changes produces a format perfectly readable by Le Progeciel and by all other statistical or phylogenetic programs that display objects (strains) in rows and descriptors in columns.

Preparation of comprehensive matrices

Characterization of strains is a normally achieved by studying their DNA with several molecular procedures as RFLP, AFLP, RAPD, other PCR applications and electrokaryotyping, which can be considered separately or all together since they are different and independent descriptors of the same strains. When many strains are studied with different techniques, investigators might therefore face the problem of analyzing together all the data from these different techniques, as well as performing an analysis with the data from a single technique.

Let's consider these two problems separately with the following example: we are characterizing 64 strains with three different RAPD primers and with electrokaryotyping (EK), although our agarose gel can accommodate no more than 18 samples. The first problem is that analyzing all the 64 samples will require four independent agarose gels that will be reported as four different images. The procedures outlined above allow analyzing all these data simultaneously, by producing a single column of molecular weight data to use a single input for classification (see processing of molecular weight data from Mac Curve Fit). The second problem, that of carrying out a single analysis with all the data from the three RAPDs and the EK, can be solved by the following procedure: calculate all four binary matrices and copy each one side by side in a new Excel file (see saving format of binary matrices). This will produce a comprehensive matrix with as many rows as strains (64) and as many columns as the sum of the classes obtained with all four binary matrices. In our case, if the first RAPD yielded 10 classes, the second 22, the third 19 and the EK 18 we will obtain a 64 by 69 matrix. This operation can be preceded by a Mantel test as described below.

The binary classification system is crucial to solving both the problems outlined above. In fact using more matrices with quantitative molecular weight data will produce problems outlined below. Combined matrices may display many missing values caused by the accumulation of missing values of each single matrix. (Tab. **XX**). Moreover, there is the probability of having very

similar (or even identical) values of molecular weight from different single matrices, which will produce duplicated quantitative data for the same object, coming from different techniques. On the other hand, the use of migration distance values (as in some commercial packages) precludes both an effective analysis of gels carrying DNA from the same procedure and the cumulative analysis of data from different techniques.

Table XX: Example of a matrix of quantitative data displaying seven missing values marked as m.v.

Sample A	1050	800	500	355
Sample B	750	400	**m.v.**	**m.v.**
Sample C	800	355	**m.v.**	**m.v.**
Sample D	650	**m.v.**	**m.v.**	**m.v.**

The aim of the next chapters is not to instruct the investigator in the use of statistical or phylogenetic programs, but to outline the way to use ClassMaker matrices as input in these applications and to describe some of the feasible analytical strategies. Specific details on the use of each software program can be found in their instruction manuals and in related papers.

F. Statistical analysis with "Le Progeciel"

This application is focused on statistical multivariate analysis. It uses as input **rectangular** matrices (i.e. those with descriptors in columns and cases in rows) and can calculate statistical distances or similarities among strains as distance or similarity matrices. These matrices are defined **squared** or **resemblance** matrices, because strains are reported in both columns and rows.

In resemblance matrices comparisons between identical strains are reported on the descending diagonal. In a similarity matrices these will be the highest values, while in distance matrices the lowest (0). The program has many algorithms for calculating distances or similarities and three algorithms for transforming similarity into distance matrices or *vice versa*. With rectangular matrices one can carry out a Principal Component Analysis (PCA) as shown in Fig. 2, . With distance matrices it is possible to calculate and draw a tree or dendrogram (Fig.1) or obtain a Principal Coordinate Analysis (PCoA) as showed in Fig. 3. Among other functions useful for the biologist, Le Progeciel allows comparisons among different similarity (or distance) matrices and biogeographic studies.

Let's now provide some practical details for carrying out the procedures outlined above.

A.Importing the input (rectangular) matrix

a. Excel files saved as tables can be imported with the function **import** of the **edit** menu.
b. Different distance or similarity algorithms can be applied depending on the requirement of the investigation. Refer to the Le Progeciel manual for an extensive description of the meaning of each procedure. Note that every type of similarity or distance matrix can be normalized, producing figures ranging from 0 to 1, with one of the **range and standardize** functions in the **VerNorm** submenu of the **Modules** menu. Some distance formulas already include normalization in their algorithm.
c. PCA and PCoA analyses can be carried out respectively with the **Principal components** and the **Principal coordinates** commands of the **Modules** menu.
d. Dendrograms can be calculated and drawn with the free application T-Rex available from the same web site of Le Progeciel. An alternative routine with more options will be outlined in chapter F because it requires the use of Phylip and a tree drawing application.

e. The analysis of the correlation among resemblance matrix is carried out with the **Mantel** test available from the **Modules** menu. This test is very useful when comparing two sets of descriptors of the same set of strains (for instance two RAPD primers). Let's consider again the above example of the 64 strains. One could want to test the correlation between the three similarity matrices obtained separately from the three RAPDs. Following the instructions of Le Progeciel and the explanations of the manual, one will obtain from each **Mantel** test of two matrices (e.g. similarity matrix from RAPD-one and RAPD-two) a value ranging from -1 to 1, with -1 indicating absolute but negative correlation (i.e. high values of similarity obtained with RAPD-one correspond to low values of similarity in the RAPD-two similarity matrix), +1 indicating absolute and positive correlation and 0 implying no correlation at all. There is also a function called **three-way Mantel** capable of comparing three matrices at once. Finally, Le Progeciel can be implemented with some extensions already available with the application. Other extensions can be developed in agreement with the author.

G. Statistical and phylogenetic analysis with "Phylip"

Phylip is one of the first and remains one of the most complete and popular packages for statistic and phylogenetic analysis. It can be freely downloaded from its web site where there is also a rich collection of links to the web pages of many other programs for statistic and phylogeny in biology. One of the advantages of Phylip is the that it is an open source package already interfaced for both Mac and PC.

The input of a rectangular or resemblance matrices from ClassMaker can be carried out as follows:

Input of rectangular matrices

- using an Excel worksheet, write in the cell A1 the number of the strains or of the species in the matrix (cases)
- in B1 write the number of descriptors (classes or variables)
- With the pointer in A2 copy the whole binary matrix with cases in the rows and variables in the columns
- Save as text

Input of resemblance matrices

- Write the number of strains or of species, preferably in Simple Text, then press enter
- On the second line copy the whole resemblance matrix . Figures must have not more than three digits. It is preferable to set the font to Courier or any other monotype font. Save as text.

Remember that Phylip applications read only files in the same folder as the application and that names of the input files must be entered in each Phylip application with the keyboard.
Refer to the Phylip manual for any additional explanation.

H. Combined use of Le Progeciel , Phylip and tree drawing programs

In addition to the procedure with T-Rex outlined above, dendrograms can be calculated and drawn with Phylip and a drawing program as follows:

- Set the preferences of Le Progeciel at three decimals
- Calculate the similarity matrix and copy it by clicking on the upper left square to highlight the whole matrix.
- Open a new text file in Simple Text or in another writing program
- Write the number of samples in the matrix (cases)
- In the second line copy the matrix.
- Save the file as text.
- Place the file in the same folder as the applications of Phylip.
- Calculate the dendrogram with **Neighbor** (Neighbor Joining, or UPGMA), **Fitch** or **Kitch**.
- The output will be a text file called treefile that will appear automatically in the same folder of the Phylip applications.
- Trees can be drawn with the applications of tree drawing available in Phylip or with other free programs such as NJ Plot, TreeViewer or DendroMaker DM4. In any case the treefile should be opened from one of the tree-drawing applications.

This procedure allows drawing trees using several options such as editing, branch length, swapping, definition of new outgroup etc.

I. Phylogenetic analysis with PAUP and MacClade

Although PAUP and Mac Clade are not free software, they have been included in this manual because of their effectiveness and affordable price. Since both applications use the **nexus format**, this short chapter aims to allow the use of the binary matrices from ClassMaker as input according to this procedure:

- Save the rectangular matrix (the only possible input) in Excel as text
- With Mac Clade open the Excel file as "simple table".
- If one wants to process the matrix with PAUP or other programs using the nexus format, save the file as a normal nexus file.
- With PAUP open the file using the command execute.

Other details can be found in the manual of MacClade; PAUP has not a real manual but a good on-line guide to refer to.

J. Ad majora

Ad majora is a Latin expression meaning the desire to progress. It is also the wish of ClassMaker to be of use to all investigators involved in characterization and to improve with further developments. We therefore encourage everybody interested in this application to suggest improvements, to indicate bugs or deficiencies. Everybody interested in program development can obtain the password for the code source of this Excel Macro. *Ad majora*!